Course Name: Machine Learning and Deep learning - Fundamentals and Applications

Professor Name: Prof. M. K. Bhuyan

Department Name: Electronics and Electrical Engineering

Institute Name: Indian Institute of Technology, Guwahati

Week-2

Lecture-5

Welcome to NPTEL MOOCs course on machine learning and deep learning, Fundamentation Applications. In my first class, I discussed the concept of pattern classification. In a typical pattern classification system, the first step is feature extraction. After feature extraction, I have to consider or I have to select the most discriminative features. In a particular pattern classification system, all the features may not be useful. So that is why I have to consider the most discriminative features

and that is called the feature selection.

And after feature selection, I will be getting the feature vector and based on this feature vector, I can do the classification. I can apply pattern classification algorithms, machine learning techniques for object recognition, object detection, object classification. So this is a typical structure of a pattern classification system. Today in this class, I will discuss the same concept, the concept of pattern classification and

mainly I will consider the concept of the statistical machine learning.

So let us start this class, what is the statistical machine learning techniques. So in my first class, I discuss the concept of the pattern classification. So I am repeating this. So for a typical pattern classification systems, suppose I have some patterns and sensor is available for taking the measurements and after this what I have to do, I have to consider the feature generation that is the feature extraction, feature generation and after the feature generation, what I have to consider, I have to select the most discriminative features.

This is called the feature selection.

And finally based on these features, I have to consider the classifier design. So I can consider the classifier for decision making for pattern classification, object recognition like this and maybe I can consider the system evaluation. So that means to evaluate the performance of the systems, I am considering this step that is the system evaluation. So there may be the feedback from the system evaluation step.

So the feedback may be something like this.

So you can see I am giving the feedback from the system evaluation so that I can improve the performance of the systems. So like this, there may be some feedback and based on this feedback, I can improve the performance of the systems. So this is a typical pattern classification system. The same thing I can show like this also. Suppose I have the patterns and suppose I am doing the measurement by using sensors and I am having the measured values.

So this is nothing but feature extraction. After this, I am getting the feature values and if I apply the feature selection techniques, so I will be getting the feature vector. So the feature vector is x and based on this feature vector, I can do the classification. So the feature vector is nothing but the mathematical description of a particular pattern. So this is the d dimensional feature vector and based on this feature vector, I can do the classification.

So classification may be something like this. So input is the feature vector and we are considering the classifier and we have a database. This is database or maybe we can consider information. Some information is available or maybe some rules are available and based on these rules, based on this information, I can do the classification. So in a supervised learning techniques, for each and every classes, I have the training samples and based on these training samples, I have to train the classifier.

So this is called the learning and after this, I am getting the learned model and based on this learned model, I can do the classification. So you can see I am showing the information and the rules. That means I have some rules for classification and based on this rule, I am doing the classifications. Classification means the decision making and there may be two types of decision making. One is the hard decision making and another may be soft decisions.

In hard decision, we consider classical set theory and in case of the soft decision, we consider fuzzy logic. So what is the hard decision and what is the soft decision? Suppose I have some samples, these are the samples belonging to the class omega 1 suppose and these are some samples belonging to another class. Suppose the class is omega 2. This is the discrete decision boundary. Suppose if I consider, suppose this sample if I consider, there is no possibility that this particular sample may belong to another class.

The decision boundary is fixed between the classes and there is no possibility that a particular sample may belong to another class. But in case of the fuzzy logic, the decision boundary is not like this. Again, I am considering the same thing. So these are the samples belonging to one class and these are the samples belonging to another class.

So the decision boundary is not rigid.

So in this case, there is a possibility that a particular sample may belongs to this class also. There is a possibility. So there is a possibility that this sample may belong to another class. This possibility is defined by membership grade. So in the fuzzy logic, there is a concept.

This is the membership grade and it lies between 0 and 1. So based on this membership grade, I can take a decision that means there is a possibility that a particular sample may belong to another class. So this concept I will explain later on what is the fuzzy logic and what is the soft decision. So in my future discussions, I will explain what is the soft decision and in my class mainly I will consider the hard decision making.

So let us move to the next slide.

So for a pattern classification system, we have to consider most discriminative features. So suppose I have the samples like this. For a particular class, the class is suppose omega 1 and these are the samples belonging to another class. This is class omega 2.

So this is the decision boundary.

It is easy to draw a decision boundary between the classes. So I can say this is the good features because I can easily draw the decision boundary between the classes. But suppose the samples are like this. So suppose these are the samples.

Then in this case, it is very difficult to draw the decision boundary.

So maybe I can consider them as bad features. So for a particular pattern classification system, we have to identify the most discriminative features and this is called feature selection. And in my first class, I have shown one mapping diagram, the same thing I am showing here. Suppose I have these classes, the class omega i, class omega j and another class is class omega k, three classes we are considering. So in my first class, I explained this and corresponding to these classes, I have the patterns.

So this is P4, suppose P1, these are the patterns and suppose the pattern is P2, another pattern is P3, like this I have the patterns. So this space is called the class membership space. This is called pattern space. And suppose I am considering another space that is the measurement space. So I have three measurements M1, M2, M3 suppose.

And let us consider the mapping. So corresponding to the class omega i, I have two patterns, pattern P1 and P4. This is the mapping suppose gi is the mapping and corresponding to the class omega j, the mapping is like this. So this mapping is suppose gj. So corresponding to the class omega j, I have a pattern, the pattern is P2 and corresponding to the class omega k, the pattern is suppose the P3, I have only one pattern, the mapping is gk.

So corresponding to the class omega i, I have two patterns P1 and P4. And after this I am showing the mapping from the pattern space to the measurement space. So corresponding to P1, the measurement is M1 suppose. Corresponding to P2, the measurement is suppose M2. Corresponding to P3, the measurement is again suppose M1.

And corresponding to P4, the measurement is M3. So here you can see that it is a mapping from the class to the pattern space and from the pattern space to the measurement space. And here you can see it is not the one to one mapping. Now what is the pattern classification? What is the definition of pattern classification? From the given measurement, I have to determine the corresponding class. That means it is the invert mapping.

Invert mapping is from the measurement how to determine the corresponding class. But the problem is it is not one to one mapping. So had it been one to one mapping, the pattern classification problem would have been very easy. But unfortunately, it is not one to one mapping. And the problem is from the measurement I have to identify the corresponding class.

And in this figure you can see there may be some overlapping here you can see this is the overlapping. This is the overlapping. So what is the meaning of this overlapping? This is overlapping because the patterns from different classes may share some common attributes. So that is why there may be some overlapping. I am repeating this the patterns from different classes may share some common attributes.

So that is the overlapping is taking place in the pattern space. So statistically what is the meaning of the pattern classification? Statistically we have to determine the probability of Omega J. Omega J means it is a class and X is the Feature vector. We have to determine this. So that means what is the probability of obtaining a particular class given the Feature vector.

So this probability we have to determine what is the probability of obtaining a particular class given the Feature vector that is the objective of the statistical pattern classification.

So let us move to the next slide. Again for the statistical pattern classification, we consider one function and that function is very popular. So the name of this function is the discriminant function. So this function is used for classification.

So it is represented like this gi x. This is used to partition r to the power d space the d dimensional space and here i is equal to I am considering c number of classes 1 to up to c number of classes we are considering. So discriminant function gi x is considered for classification and it is used to partition r to the power d space. So what is the decision rule to assign the Feature vector x to the class Ω_M. So what is the decision rule? The decision rule is I have to assign the Feature vector to the particular class if $g_M(x)$ is greater than gi (x) for all i is equal to 1 …., c and i is not equal to m. So that means based on the discriminant function I can do the classification the Feature vector x is assigned to the class Omega M if gm (x) is greater than gi (x) and what is the equation of the decision boundary the equation of the decision boundary is gk (x) suppose is equal to suppose gL (x).

So this is the equation of decision boundary. So I can give this example suppose I am considering this is the Feature space and this is the decision boundary and this is the region is R1 another region is R2 this region corresponds to the class suppose Omega 1 and this region corresponds to the class the class is Omega 2. So the equation of the decision boundary is g1 is equal to g2. So corresponding to this region I have to determine the discriminant function the discriminant function is g1(x) and corresponding to this region I have a discriminant function the discriminant function is g2(x).

So based on this discriminant function I can do the classification and g1 is equal to g2 that is the equation of the decision boundary.

So I will explain the concept of the discriminant function later on there are many concepts on discriminant function. So for the time being you should understand that in statistical pattern classification we consider discriminant function for classification and you can see we can also determine the decision boundary between the classes based on the discriminant function. And typically later on I will show the equation of the linear discriminant function is represented like this I will explain later on but for the time being you just understand this. And this W is the weight vector this is not the class this is the weight vector x is the feature vector and W 0 i this is the weight not the feature vector it is a d cross 1 vector this is called the weight vector and this is called the bias. So this concept I will explain later on but this is the expression for linear discriminant function.

Now I will move to the concept of the statistical machine learning that is the statistical pattern classification. In statistical pattern classification the main concept is coming from the Bayes law. So you know in the probability what is the Bayes law and statistical pattern classification technique are derived from this concept the Bayes law. So what is the Bayes decision making I will explain in my next slide. So what is the classification how I can do the classification based on this Bayes decision theory.

So in my previous classes I discussed about the concept of the regression. Regression means the fitting of a line between the sample points. But in the classification is different from regression because output will be a discrete level denoting the entity of the class. So output will be the classes maybe the class omega 1 class omega 2 so that will be the output. But in the regression what we are considering the fitting of a line or

fitting of a curve between the sample points.

So mainly we are considering the probabilistic interference. So we are considering the probabilistic theory for decision making and the Bayes theorem is the powerful tool for decision making. So in this class I will explain what is the Bayes theorem and how it can be applied for decision making. So suppose in this example if I consider a pattern classification system I have to do the pattern classification. So two classes two type of fish one is sea bass and another one is salmon.

So I have to classify these two types of fish. So the problem is need to recognize the type of fish. So it is a two class problem. So I can define the classes like this omega 1 that represents the fish salmon and omega 2 represents seabass. Now suppose if I consider that suppose the prior information is available that is the probability of omega 1 and the probability of omega 2 is available.

So based on this probability I can consider like this. So if the probability of omega 1 is greater than probability of omega 2 that means the class is salmon the class is omega 1. So based on this prior probability we can take a classification decision. That means we have to determine the probability the probability of omega 1 we have to determine the probability omega 2. And suppose in this example the probability of omega 1 is greater than probability of omega 2. Then what I have to consider I have to consider a class of omega 1 that corresponds to the fish salmon.

So suppose in this case we are considering 8000 fish and out of this 8000 fish 6000 are salmon and 2000 sea bass. So from this we can determine the probability that is the prior probability we can determine. So prior probability of salmon is 0.

75 and the prior probability of sea bass it is 0.25. And based on this prior probability we can take a classification decision since the probability of omega 1 is greater than probability of omega 2 then I can say the class is omega 1. So assign unknown fish as

salmon omega 1 if the probability of omega 1 is greater than probability of omega 2. Otherwise we have to consider the other class the other class is omega 2 and that is nothing but sea bass. So this is mainly by considering the prior probability.

So this is a simple decision making based on the probability. Now decision rule is not a good rule because it is a flawed rule because salmon will always be favored and assigned to a test fish. So because in this case we are only considering the prior information so the classification is not accurate. So we have to consider some additional information for accurate decision making. So prior probability we will be considering and some other information we will be considering.

So maybe we can consider some other Features that is the Feature extraction.

Maybe we can consider the length of the fish as a Feature the width of the fish as a Feature the color of the fish we can consider the life span we can consider texture we can consider. So all these Features we can consider and based on these Features we can do accurate classification. So we are considering number of important Features and based on these Features we can do the classification and this is the solution. So for this we have d Features and we have a Feature vector that is the d dimensional Feature vector I am getting and so x is a d dimensional Feature vector.

So this is the x 1 x 2 so up to x d so these are transpose so d dimensional Feature vector I have.

So we are considering number of Features for classification. Now what we have to consider we are considering some additional information. So what is the additional information for each class category of fish we can associate the d Features to come from a probability distribution function and that is actually nothing but the class conditional density. So what is the class conditional density I will explain later on. And the Features we are considered as continuous Features. So we have d dimensional Feature vector and we are considering the continuous Features and now we are considering the class conditional density that is the additional information

and based on this additional information I have to do the classification.

So how to improve the decision making by considering single Features. So let us work on how to improve our decision process by incorporating a single Feature x. Later on we can extend the framework for d dimensional Features. So that means we are now considering only one Feature and I am developing the decision theory how to do the classification based on this single Feature. After this the same theory can be applicable if I increase the number of Features and in this example we are considering only two classes.

At present I am considering only one Feature and in Feature we can consider more number of Features. Now this is the example of class conditional density the class conditional density is the probability of x, x is the Feature vector for a particular class omega is a particular class that is called the class conditional density. This is also called the likelihood. So for two classes you can see the distribution one is omega 1 another one is omega 2 and

I am plotting the class conditional density with respect to the Feature vector x is the Feature.

So this is the class conditional density plot. Now let us come to the Bayes theorem. So this is the Bayes theorem. In the Bayes theorem you can see I can represent like this the probability of omega j, omega j is a class given the Feature vector and that is called the posterior probability is equal to probability of x given omega j into probability of omega j divided by p x.

So I can write like this is a posterior probability likelihood into prior divided by evidence.

So if I consider two class problem so j is equal to 1 to 2. So the evidence can be written like this the p x can be written like this. So this evidence actually it has no role in classification. It is simply a normalizing factor.

So this Bayes formula I can show like this. So this is the Bayes formula. So I have the prior probability I have the class conditional density that is the likelihood and the evidence is this and I am getting the posterior probability. The posterior probability is this posterior density or the posterior probability I am getting this one. So this is the Bayes formula. So if you see this equation so with the help of these two information one is the prior information another one is the likelihood we can determine the posterior.

So how to take a classification decision. So in this plot we are showing the posterior probability plot for two classes. So we are plotting this probability of omega i given x that the posterior probability we are plotting with respect to the feature vector x is the feature vector. Now how to take a decision. The decision you can take like this suppose for two classes you can see suppose this is the Bayes formula omega j given x is equal to probability x given omega j probability of omega j and the evidence.

So already I have explained the evidence has no role in classification. It is simply a normalizing factor. So what is P x P x is nothing but if I consider two classes j is equal to 1 to 2 probability x omega j probability of omega j. So this is nothing but the scale factor. Now how to select a particular class

Choose $\Omega 1$ if $P(x \mid \Omega 1) \times P(\Omega 1) > P(x \mid \Omega 2) \times P(\Omega 2)P(x \mid \Omega 1) \times P(\Omega 1) > P(x \mid \Omega 2) \times P(\Omega 2)$

Choose $\Omega 2$ if $P(x \mid \Omega 2) \times P(\Omega 2) > P(x \mid \Omega 1) \times P(\Omega 1)P(x \mid \Omega 2) \times P(\Omega 2) > P(x \mid \Omega 1) \times P(\Omega 1)$

So just you can see we are considering two information one is the likelihood and also we are considering the prior. So based on these two information I am doing the classification. So classification actually I am doing like this. So suppose x is my Feature vector that is my input and what actually I am computing I am computing this probability of omega 1 probability of x omega 1. I am computing like this probability of omega 2 probability of x omega 2 like this for C number of classes we are determining like this probability of omega C and probability of x omega C and from this we can determine the probability of omega given x we can determine that is the posterior probability of omega 2 given x and we can also determine the probability of omega C given x.

Out of all this we have to pick the largest. So based on this I can do the classification. Now for this decision making what we can consider we can consider one measure and that is the probability of error. So how to define the probability of error I am explaining in my next slide. So what is the probability of error and based on the probability of error you can decide or you can do the classification. So what is the probability of error for a particular x for a particular Feature vector the probability of error I can define like this probability of error given x is equal to probability of omega 1 given x if we decide omega 2 and it is equal to probability of omega 2 given x if we decide omega 1.

So for minimization of this error for minimization of the error what I have to consider decide omega 1 there is a class is omega 1 if the probability of omega 1 given x is greater than probability of omega 2 given x otherwise you have to select omega 2 the second class. So you can see based on the probability of error we can decide a particular class. So probability of error for a Feature vector is written like this the probability of error given x that is equal to the probability of omega 1 given x if I consider or if I select the class omega 2 and it is equal to probability of omega 2 given x if I consider the class omega 1 and I have to minimize the error. So I have to decide omega 1 if the probability of omega 1 given x is greater than probability of omega 2 given x otherwise we have to select omega 2.

So what is the average probability of error average probability of error. So average probability of error is minus infinity to plus infinity probability of error given x Px dx. So this is the average probability of error that is equal to probability of omega 1 given x if we are considering x belongs to the class omega 2 or otherwise it will be probability of omega 2 given x if I consider x belongs to class omega 1. So this average probability of error we can determine like this. So for the error minimization what we have to do x is assigned to this particular class error minimization for error minimization x is assigned to the particular class if this probability omega j x is greater than probability of omega i given x and this is for i is not equal to j. So for error minimization this Feature vector x is assigned to the class omega j if this condition is satisfied.

So this is the concept of the probability of error and I can plot the probability of error like this. So and this is I am showing the class conditional density or the likelihood I am showing omega i and this is the Feature vector is x. So this is your curve corresponding to probability of x given omega 1. I am considering two classes and this is the probability of error x given omega 2. So this area is a common area that area I can consider as area of probability of error area of probability of error.

So I can plot this like the probability of x given omega 1 I can plot and probability of x given omega 2 I can plot and you can see I have a common area and that area is nothing but the area of probability of error. So this concept I am explaining in my next slide same thing I am repeating. So here you can see how to take a decision. So if I consider the probability of omega 1 x is greater than probability of omega 2 x then I have to consider the class omega 1. So already I have explained and if the probability of omega 1 given x is less than probability of omega 2 x then I have to consider the class omega 2.

Now this probability of error given x already I have defined. So that is equal to probability of omega 1 given x if I consider the class omega 2 that already I have explained and the probability of error given x is equal to probability of omega 2 given x if I consider the class omega 1. So this is the definition of the probability of error given theFeature vector. TheFeature vector is x. So how to minimize the probability of error? So we have to select or we have to decide the class omega 1 if this probability the probability of omega 1 given x is greater than probability of omega 2 given x otherwise we have to consider omega 2. Therefore the probability of error given x we have to consider the minimum error and that is nothing but the bayes decision.

So based on the probability of error we can take a classification decision. So this is the average probability of error. So already I have explained. So this average probability of error you can determine like this. So it is from minus infinity to plus infinity because I am considering the continuous x so that I can determine the average probability of error and this average probability of error should be small as far as possible for every value of x and this is the objective of the bayes classifier.

So we have to reduce the probability of error. Now if I consider c number of classes the same theory can be extended. Suppose we have an unknown pattern and now you are considering the d dimensional Feature vector. Previously we consider only one dimensional Feature vector but suppose if I consider the d dimensional Feature vector d dimensional Feature vector already I have shown. So this is a Feature vector and I have d number of Features. So this is a d dimensional Feature vector and again I am applying this Bayes rule for classification and we can compute the posterior probability of the pattern with respect to c number of classes.

For c number of classes we have to determine this posterior probability and based on this we have to select which one is the greatest posterior probability and based on this I can take a classification decision. So in the decision making we are assigning a particular pattern to the class for which the posterior probability is the greatest. So for c number of classes I have to determine this the posterior probability I have to determine and based on this posterior probability I can decide. So I have to pick the largest posterior probability and that corresponds to the particular class. So here you can see the test sample we are considering omega test and based on this we are determining the posterior probability and based on the posterior probability I am taking a classification decision and this is the evidence already I have mentioned evidence is nothing but the normalizing factor and this is same for all the classes.

So it has no role in classifications. So how to select a particular class the largest posterior probability we have to consider and that corresponding class we have to determine that is the decision making we have to do. So the pattern is assigned to this class that class corresponding to the largest posterior probability and there is another definition for decision making another parameter I can say that is the Risks. So in the Bayes theorem you see this is omega J x is equal to P x omega J probability of omega J. So this is the Bayes theorem already I have explained and this is the normalizing factor or the evidence is for c number of classes J is equal to 1 to c.

So we have to determine this x omega J P omega J. Now how to define the Ricks suppose for a particular x, x is the feature vector so x may belongs to the class omega 1 like this or it may belongs to class omega 2. So it may belongs to class omega c like this we can consider. So that means it may belongs to the class omega 1 that means x is assigned to the class omega 1 or maybe we can consider is x may be assigned to the class omega 2 like this. So suppose I am considering x is assigned to the class omega 1 so for this some action is taken some action or maybe the decision is suppose alpha i is taken.

So the loss is defined like this the loss is nothing but lambda alpha i omega J. So what is the definition of the loss I am taking the x on the x 1 is alpha i corresponding to the class the class is omega J. So some action is taken for the class omega J. So loss is defined like this lambda alpha i given omega J the loss is defined like this lambda i J. So suppose I have the classes omega 1 omega 2 omega k like this these are the classes and I am taking some actions the x 1 is alpha 0 alpha 1 like this suppose alpha k dash this actions I am taking. So suppose if I take this action so corresponding to the class omega 1 the x 1 alpha k dash we are considering.

So what will be the loss the loss will be lambda k dash 1. So as per the definition of the loss the loss will be lambda k dash 1 this is the loss. So loss is defined like this. So now from the loss I will define the risk so how to define the risk so that I will explain in my next class. So I have defined what is the loss the loss is defined like this so x belongs to a particular class omega 1 and for this I am taking some actions and the action is alpha i. So the corresponding this the loss is lambda alpha i given omega J and corresponding to this I can determine the loss the loss is lambda i J and based on this loss I can determine the risk.

So in my next class I will explain what is the risk. So let me stop here today in today's class so briefly I have explained the concept of the Bayesian decision making. So how to take a decision based on these probabilities we are considering the posterior probabilities this posterior probability we can determine from the likelihood and also from the prior probability and the evidence has no role in classification because it is a simply a normalizing factor. After this I discuss the concept of the probability of error. So with the help of the probability of error you can take a classification decision.

After this I define for a particular action what is the loss. After defining the loss in the next class I will discuss the concept of the risk. So with the help of this parameter the risk I can do the classification I can take a classification decision. So in my next class I will discuss about these concepts. So let me stop here today. Thank you.